基于 Cassandra 的海量 MUSER 数据分布式存储与检索研究*

石玥[1]，王锋[1,2]，李鹏程[1]，刘应波[1*]
(1.昆明理工大学云南省计算机技术应用重点实验室，云南 昆明 650504
2.中国科学院云南天文台，云南 昆明 650216)

摘 要：明安图射电频谱日像仪每天产生海量观测数据，传统的关系型数据库存储和管理这些数据时，面临着读写延迟高、性能和容量扩展能力有限以及可用性弱等诸多问题针对这些问题，开展了基于 NoSQL 的海量数据存储与检索应用研究。首先，详细分析了明安图射电频谱日像仪数据特点、存储需要以及面临的问题；然后，对明安图射电频谱日像仪进行数据建模，给出了明安图射电频谱日像仪的列式非关系型数据模型，同时提出了元数据和数据在 NoSQL 中的同步存储方法，解决了二者的一致性问题，在此基础上实现了基于 Cassandra 的海量天文数据存储管理系统(MBDMS)，最后通过实验验证了系统存储与检索的高效性、扩展性以及可行性，从结果来看，MBDMS 可以很好的满足数据管理需要，是解决当前数据存储问题的一种有效方案。

关键词：非关系型数据库； 存储； 检索；
中图分类号：TN216　文献标识码：A　文章编号：1672-7673(2018)03

明安图射电频谱日像仪(Mingantu Ultrawide SpEctral Radioheliograph,MUSER)是同时以高时间、高空间和高频率分辨率对太阳进行射电频谱成像的设备，主要对日冕层进行层析观测，探测日冕大气，研究太阳活动的动力学性质。项目分为两期：第 1 期低频阵（MUSER-I）由 40 面 4.5m 口径的抛物面天线及接收设备组成，在 64 个频率点上成像；第 2 期高频阵（MUSER-II）由 60 面 2m 口径的抛物面天线及接收设备组成，可以在 528 个频率点上成像[1]。

明安图射电频谱日像仪的观测数据需实时存储并处理以研究太阳动向和监控天线观测状态。对目前基于关系型数据库的数据处理具有较大的技术挑战。明安图射电频谱日像仪数据的存储需求具体如下[2]：

（1）数据存储。需要能够支持每月低频阵和高频阵约 32TB 和 70TB 数据的存储工作。

（2）数据的一致性存储。需要把帧头数据和帧数据处理后的图像进行统一存储。

（3）数据的管理和检索。需要实现明安图射电频谱日像仪数据管理系统，更直观的把数据库中存储的数据呈现出来，在完成图像相似匹配的前提下实现信息的高效检索。

在现有存储方案中，数据的一致性不能得到有效保障，在数据量大的情况下，很容易发生帧头信息和帧数据不匹配的问题。非关系型数据库的出现，为高速同步存储海量帧头数据和帧数据提供了可能，从而为一致性问题的解决提供了新思路。

1 国内外研究现状

海量数据的存储管理是天文领域的一个重要问题，随着天文大数据的产生，各个天文研究机构也开始研究大数据存储管理方案。美国华盛顿大学天文系的计算机科学家采用 HDFS 分布处理海量天文图像数据，使用 MapReduce 将天文图

* 基金项目：国家自然科学基金(61462053, 11703010)；中国博士后科学基金(2016M602730) 资助.
收稿日期：2017-11-27；　修订日期：2018-01-05
作者简介：石玥，女，硕士.研究方向：分布式计算.E-mail: 412630885@qq.com
通讯作者：刘应波，男，博士.研究方向：分布式计算、海量数据存储与检索. E-mail: liuyingbo@astrolab.cn

像数据按组分解成小型文件序列后再输入系统，在减少文件总量的情况下明显提高处理效率。文[3]提出了一种针对海量数据的新型数据管理技术-负数据库，利用观测数据的补充集来获取必要的信息，从而实现对明安图射电频谱日像仪数据的高效管理。文[4]提出了使用 NoSQL 对 FITS 文件头元数据进行存储研究，并且对其可行性进行了实验论证。为了满足海量天文数据的高性能检索和查询需求，文[5]提出了一种基于 ElasticSearch 分布式搜索引擎，实现了海量 FITS 数据高效检索的方法。文[6]提出了一个基于 Cassandra 的分布式反向索引，用以解决传统关系型数据存储无法解决的可扩展性问题，并在此基础上设计了数据模型和查询处理过程。理论上这个方法同样可以用在天文数据处理中。这些工作虽在一定程度上提升了工作效率，但是仍在数据高效存储方面有所不足。

文[7]实现了一个基于 NoSQL 的高性能存储系统，对数据的存储位置和数据查询结构进行深入研究，保证了数据存储的灵活性、可移植性和稳定性。文[8]针对传统关系型数据库无法满足海量数据的存储与访问需求的问题，提出了基于 NoSQL 的分布式存储与扩展解决办法，提出将 NoSQL 作为镜像引入数据库架构系统，在一定程度上避免了资源浪费以及服务器过载。经相关实验论证，上述两种研究均可应用于天文数据处理方面。

文[9]基于 NoSQL 技术设计了明安图射电频谱日像仪数据归档与发布系统，使用 FastBit 数据库对数据进行存储研究，利用位图索引的优势，大大提高了索引查询的效率。对海量数据存储和检索的有效手段的研究和探讨与本文的研究存在较大的不同，前者偏重于数据检索效率,本文偏重于数据存储的可靠性、可用性以及一致性。

2 数据一致性问题

2.1 一致性问题产生的原因

元数据通常用来描述数据，例如明安图射电频谱日像仪采集图像的时间、极化方式等。因此元数据成了支持数据检索的关键，而数据检索功能是数据管理的重要部分。通常这类数据的存储方式是元数据和数据文件分离存储，这种方式带来了数据的一致性问题，元数据或数据一旦其中之一出现丢失，数据之间就存在不匹配的情况，特别是在如此巨大的数据背景下，这种丢失更容易发生，因此在这种情况下，元数据和数据之间的一致性关系需要得到保证。为了方便后续分析阐述，对 MUSER 数据的一致性问题进行如下描述。

定义：为数据，为的元数据，由多个字段 属性构成。用集合关系可表示为

. (1)

二者之间的一致性关系定义为,不一致关系为，而不一致性关系又细分为(1),存在数据，但是无法找到对应的元数据；(2),存在元数据但是无法找到对应的数据；(3)，二者均不存在。对于这些处理方法参见文[10]。

2.2 传统存储过程中的解决方案

对于一致性解决方案有如下两种：

(1)简单的一致性处理方案。使用异步非阻塞的方式存储元数据和数据文件，二者不存在一致性的协商手段，主机的可靠性是保证数据一致性的前提，如图 1。

(2)两段提交协议。文[10]使用成熟的两段提交协议做数据服务之间的同步，

文中把FITS文件各个参与部分按照两段提交的角色进行划分，如图2。数据采集服务器充当协调者的角色，元数据和数据充当成员的角色，三者直接通过两段提交协议进行一致性确认。

图1 一致性问题处理
Fig.1 Consistency problem processing

图2 两段提交协议
Fig.2 2PC

2.3 基于NoSQL的一致性解决方法

数据和元数据同步存储，在数据量小，性能要求不高的场合，可以很容易处理。针对明安图射电频谱日像仪的UVFITS文件，许多关系型数据库都提供了大数据的存储方式，利用MySQL的Longlob，PostgreSQL的Bytea，SQLServer的Blob，Oracle的Blob和Clob等，但这种方式存储的最大问题在于关系型数据库的原子性（Atomicity）、一致性（Consistency）、隔离性（Isolation）、持久性（Durability）的特性限制，导致了在面对海量天文数据的时候出现的写入延迟高、水平扩展能力差以及数据结构固定等诸多问题。

针对这些问题，本文研究基于NoSQL的分布式海量数据存储方案，通过调研和研究，选取Cassandra[11]作为底层数据存储平台，相对于传统的关系型数据库，Cassandra具有存储速度快、扩展性高、数据结构随机等特点，此外与其他NoSQL数据库相比，还具备以下优点：

(1)无中心架构，单点故障不会造成系统运行中断。

(2)动态水平扩展，新节点的加入不会影响当前工作的进程。

(3)存储模式具有更高的灵活性，可以在系统运行时随意为记录添加或删除字段。

(4)高并发读写能力，超级列和列族概念的引入，使得键值匹配次数减少，减少了文件数据的寻址时间，可以实现高速读写数据。

新数据模型采用数据文件和元数据在一起的方式进行数据存储，这种方式不仅避免了一致性问题，同时在性能上也较传统的方式有很大的提高，表1为Cassandra数据库和MySQL数据库插入数据情况对比。相比于MySQL采用的共享内存的存储机制，Cassandra具有的高并发读写能力以及无中心架构更加有利于保证数据存储的高效性以及一致性。

表1 性能对比

Table 1 performance comparison

	Number(thousand)	Size(GB)	Time(min)	Rate(GB)
MySQL	5	1	3.5	0.28
	10	2	7.1	0.28
	20	4	13.9	0.29
Cassandra	5	1	2.1	0.47
	10	2	3.8	0.53
	20	4	6.9	0.58

3. 基于NoSQL的海量天文数据存储系统MBDMS

3.1 明安图射电频谱日像仪数据建模

明安图射电频谱日像仪可设置在循环和非循环两种模式下进行观测。循环

模式即天线的观测在各个射频频段之间循环进行，非循环模式即天线的观测频率固定在同一频率范围。明安图射电频谱日像仪有两种极化方式，左旋和右旋；有四个波段，分别为0.4~0.8GHz、0.8~1.2GHz、1.2~1.6GHz、1.6~2.0GHz。每个波段，有16个通道，每个通道的带宽为25MHz。在实际观测中，每3ms生成一帧数据，每帧数据由帧头和数据组成，总数据量为0.1MByte。每一分钟可以产生19 200帧数据。研究针对于明安图射电频谱日像仪数据的观测时间、文件名、极化方式、种类、频率等信息进行存储。其中观测时间(time)为主键，设置为非空。filename为文件名及存储路径，设置为非空。极化(polarization)和频率(frequency)等均为帧头的参数，result为帧数据转换成的图片文件，如图3、图4。

Cassandra可以看成4维的哈希结构构成的Key/Value数据模式。其包括键值空间(Keyspace)、列族(Column Family)、键值(Key)和列(Column)，以三级嵌套的形式存在。如图3和图4，为Cassandra数据库的两种数据存储模型设计。图3的设计模式为每一个CF对应一个Column，图4设计模式为每一个CF对应多个Column。在读写操作中，第2种设计模式可以减少CF对key值的匹配，减少文件寻址时间，从而减少系统开销。本文的数据模型采用第2种设计模式。Cassandra数据库存储数据的流程为提交动作记录到日志，然后把数据写入内存Memtable中，等达到系统设定条件，再将Memtable中的数据批量写入磁盘，存储为SStable结构。

图3 数据存储模式1
Fig.3 The first data storage mode

图4 数据存储模式2
Fig.4 The second data storage mode

3.2 MBDMS实现

图5为系统层次结构图。Client端接受用户的请求，并对用户做一些合法性检查，把用户的请求发给服务器端。服务器端接收到用户的请求后，对用户的请求做出解析，处理用户的请求，并把操作数据的命令返回Data端，Data端进行相应的操作。服务器端通过Data的返回信息判断用户的操作是否成功，返回信息给Client端。其中客户端目前需要3个关键的功能模块。第1个模块为数据处理模块，在这个模块中，用户可以查询和删除数据库中的数据信息，并且可以实现分页显示效果；图像转化模块中用户可以根据实际需要，把选定的数据中的图像数据转换成图像，以便于用户查阅；图像检索模块中用户可以根据自己的需求，对提交的图像进行特征值对比检索，用以筛选出数据库中相似度最高的图像。论文基于两种数据库接口，利用Python和Django框架实现了海量天文数据存储系统。明安图射电频谱日像仪数据展示如图6。

面对明安图射电频谱日像仪数据海量、非结构化、一致性弱等特性，客户端的用户主要分成两类，(1)对MUSER元数据进行整理、加工的管理者，该类用户对数据侧重方向为元数据的统一性、规整性，以便于更高效地实现MUSER元数据的存储。(2)对明安图射电频谱日像仪数据进行处理的使用者，该类用户需要通过数据库中存储的元数据检索出帧数据，对元数据和帧数据的一致性有更强的要求。

图5 系统存储与检索流程图
Fig.5 The System storage and data retrieval diagram

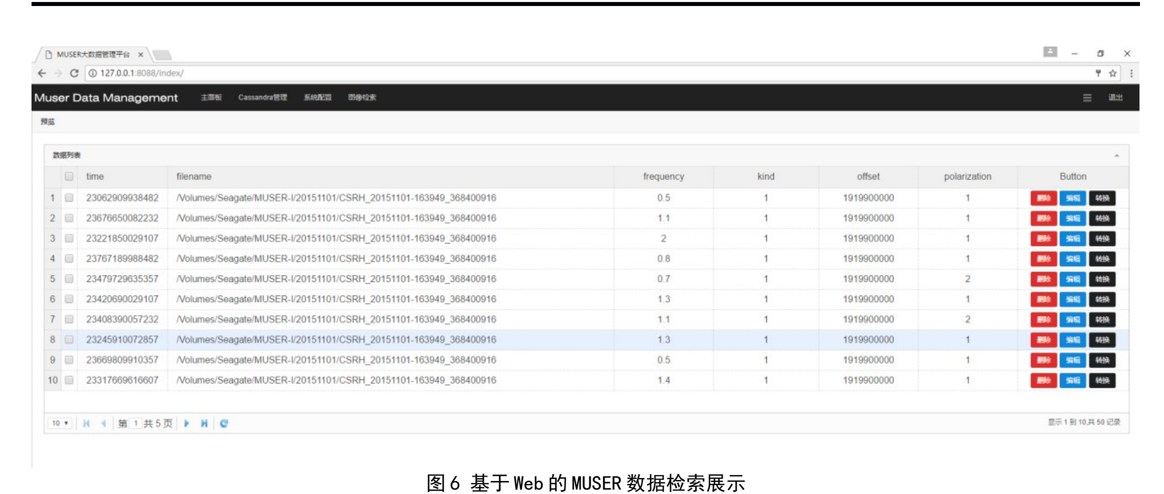


图 6 基于 Web 的 MUSER 数据检索展示
Fig.6 The exhibition of MUSER data retrieval by using Web GUI

4 性能测试

为验证 MBDMS 性能，通过其提供的关系型和非关系型数据接口连接 MySQL 和 Cassandra 数据库，在此基础上进行了 3 组测试，实验环境如下：4 台 E7200 酷睿双核，2.53GHz CPU，2G 内存，7200 转 SATA 硬盘。操作系统为 Centos 64（内核版本为 2-504.el6.x86_64）。数据库版本为 MySQL-cluster-gpl-7.2.28-linux2.6-x86_64 和 Apache-Cassandra-3.9-bin，均采用默认配置。

4.1 查询语句

实验采用如表 2 常用的检索语句。

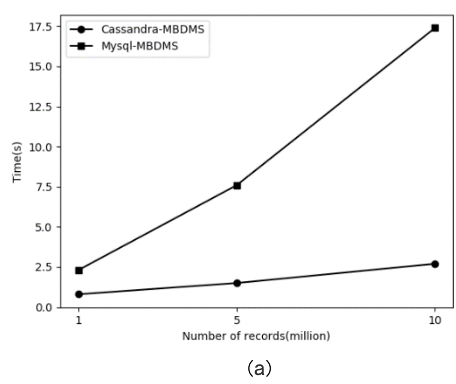
表 2 MySQL 和 Cassandra 的查询语句

Tab.2 The sentence of queried by MySQL and Cassandra

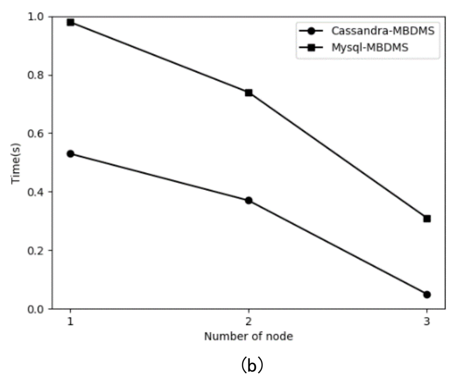
	SQL&CQL
MySQL	"select time, filename, kind, frequency, polarization, result, offset from muserdata where time=23062909938482" "insert into muserdata SET result='%s'" % \ mdb.escape_string(img)"
Cassandra	"select time, filename, kind, frequency, polarization, result, offset from muserdata where time = 23062909938482" "cf = pycassa.ColumnFamily(pool, 'muserblob') " "cf.insert(h,{'result': data})"

检索实验中使用 time(主键)为条件进行数据查询。采用元数据和帧数据统一存储的方式进行 MUSER 数据存储，保证了数据的一致性，并使用新存储策略采用的 Cassandra 与旧存储策略采用的 MySQL 数据库进行性能测试，具有一定的代表性。所需查询的数据为数据库中存有的所有数据，分别为文件存储路径、数据种类、数据频率、极化和帧数据。为了保证实验的精确性，两种数据库均使用相同的查询语句。

4.2 检索性能扩展性测试



(a)



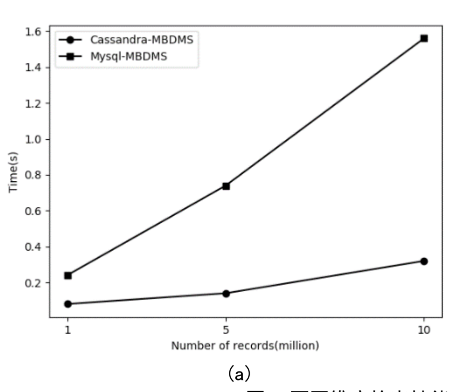
(b)

图 7 不同数据量及节点数检索性能对比，(a)不同数据量检索，(b)不同节点数检索

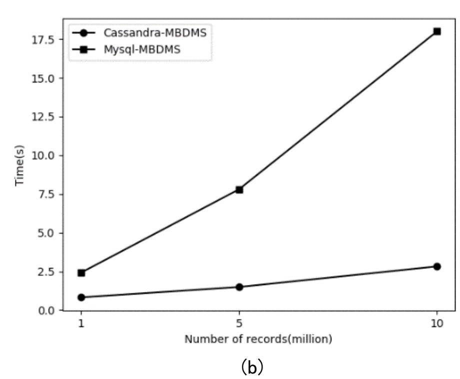
Fig.7 The comparison of data volume and number of node,(a)different data volume,(b)different number of node

图 7(a)为查询维度为 7，集群节点数为 3，当数据库中数据量不同时，两种数据库对数据进行查询的实验结果。扩展性是明安图射电频谱日像仪大数据存储的一个基本需要，随着数据量的增大，性能和存储容量也应该具有相应的扩展能力，通常数据都以在线的方式提供服务，也要求存储系统具有动态节点增加和减少的特性，从结果可以看出，基于 Cassandra 数据库的操作所需的时间要少于 MySQL 数据库。因为前者在数据插入时，不会针对一致性进行检验，而后者采用共享内存，要为一致性提供保证，因此系统开销明显大于 Cassandra 数据库。图 7(b)为数据量为 5 百万，查询维度为 7 时，当集群节点数不同时，两种数据库的查询性能对比结果图。从结果可以看出，节点数越多，集群的性能越好。通过实验可以看出，使用 Cassandra 非关系型数据接口时 MBDMS 系统的性能更优。

4.3 查询维度对比



(a)



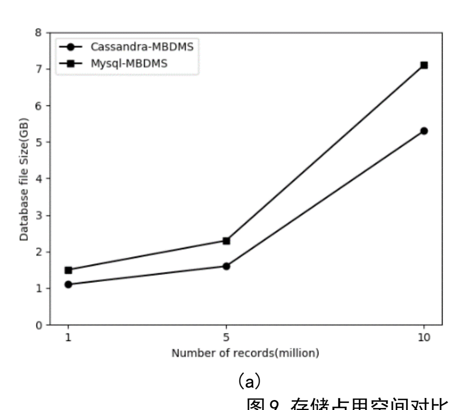
(b)

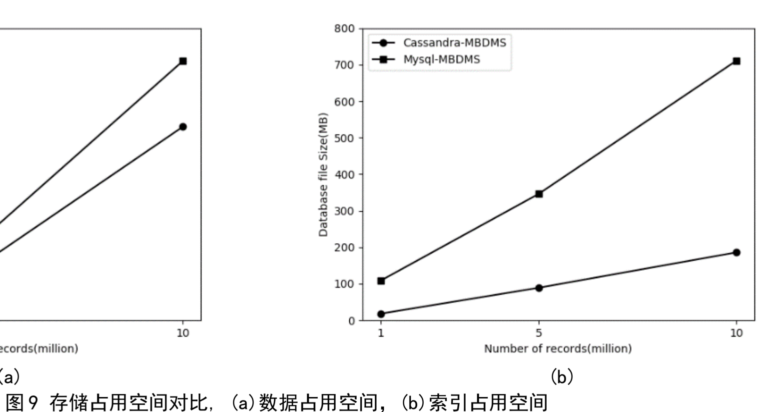
图 8 不同维度检索性能对比，(a)单维度检索，(b)多维度检索

Fig.8 The comparison of different dimensions,(a) single dimension,(b) multiple dimensions

图 8(a)和图 8(b)分别为对数据进行单维度和多维度（维度数为 7）查询时，两种数据库的性能对比情况。

由实验结果可以看出，当查询维度增加时，两种集群的系统开销均会增大，但是 MySQL 数据库所需时间增加的速率明显高于 Cassandra 数据库。因为 MySQL 数据库共享内容的存储方式与 Cassandra 数据库索引树分节点的存储方式不同，当多个进程同时查询数据时，对索引形成的压力较大，使系统开销急剧增加。





(a) (b)

图 9 存储占用空间对比，(a)数据占用空间，(b)索引占用空间

Fig.9 The comparison of storage space, (a) the space of data, (b) the space of index

4.4 索引占用空间对比

图 9 对两种数据库存储相同数据时所占用空间以及索引所占用的空间问题进行了分析。对于一个高效的海量数据存储体系而言，减少磁盘的开销是很有必要的。在测试中分析了磁盘空间占用情况，可以发现，对于 MBDMS，非关系型数据库比关系型数据库所占用空间小很多，在数据量相同的情况下，非关系型数据库对于海量数据存储更有优势。

通过上述几组实验可以看出，当数据集比较小时，例如在五百万以下的时候，MySQL 数据库的性能基本可以满足存储和检索需求。时间在 7.5s 左右。当数据集比较庞大，超过一千万的时候，可以看到，MySQL 数据库的查询时间已经超过了 17s，显然，该数据库查询的性能已经不适合再进行后续的数据处理。在上述单维度查询、多维度查询和插入的实验中可以看到，MBDMS 系统调用 Cassandra 数据库接口时，系统性能有明显的优势。

5 结束语

本文通过 Cassandra 实现了明安图射电频谱日像仪帧头数据和帧数据高速的同步存储，解决了关系型数据库环境下二者分离存储带来的一致性问题，且能够通过 Cassandra 的扩展性提供海量天文数据的适应性。同时，MBDMS 能够很好地与 Cassandra 数据库结合，满足常用的天文数据管理需要，在后续工作中，将一进步完善并优化 MDBMS 的存储和检索性能，提供其他常用主流 NoSQL 数据库数据访问接口。

参考文献

[1] 梅盈,刘东浩,王锋, 等. 中国频谱射电日像仪 FITS-IDI 文件格式研究[J]. 天文研究与技术—国家天文台台刊, 2014, 11(4)：388-395.

Mei Ying, Liu Donghao, Wang Feng, et al. A study of the FITS-IDI Format for the Chinese Spectral Radio Heliograph[J]. Astronomical Research & Techonolgy—Publications of National Astronomical Observatories of China, 2014, 11(4)：388-395.

[2] 梅盈. MUSER 海量数据预处理关键技术研究[D]. 昆明：昆明理工大学, 2015.

[3] Shi Congming, Wang Feng, Deng Hui, et al. High performance negative database for massive data management system of the Mingantu Spectral Radioheliograph[J]. Publications of the Astronomical Society of the Pacific, 2017, 129(978): 1- 12.

[4] 刘应波, 王锋, 季凯帆,等. 基于 NoSQL 的 FITS 文件头元数据存储和查询研究[J]. 计算机应用研究, 2015, 32(2):461-465.

Liu yingbo, Wangfeng, Ji kaifan, et al . Research on metadata storage and querying of FITS file based on NoSQL[J]. Application Research of Computers, 2015, 32(2):461-465.

[5] 陈亚杰, 王锋, 邓辉,等. ElasticSearch分布式搜索引擎在天文大数据检索中的应用研究[J]. 天文学报, 2016, 57(2):241-251.
Chen Yajie, Wang Feng, Deng Hui, et al. The application of ElasticSearch in the massive astronomical data retrieval[J]. Acta Astronomica Sinica, 2016, 57(2):241-251.
[6] 唐李洋, 倪志伟, 李应. 基于Cassandra的可扩展分布式反向索引的构建[J]. 计算机科学, 2011, 38(6):187-190.
Tang Liyang, Ni Zhiwei, Li Ying. Scalable Distributed Inverted Index Built on Cassandra[J]. Computer Science, 2011, 38(6):187-190.
[7] 侯朋朋. 一种高性能NoSQL存储系统的设计与实现[D]. 北京：中国科学院大学, 2013.
[8] 潘洪志. 高性能NOSQL存储系统的研究与实现[D]. 吉林：吉林大学, 2014.
[9] 林茂强. 基于NoSQL技术的MUSER数据归档与发布系统的设计与实现[D]. 昆明：昆明理工大学, 2015.
[10] 梁波, 陈腾达, 于孔林,等. 分布式实时存储环境下的FITS数据一致性研究[J]. 天文研究与技术, 2016, 13(4):489-497.
Liang Bo, Chen Tengda, Yu Konglin, et al. Research on the consistency of FITS data in the process of distributed real-time storage[J]. Astronomical Research & Techonolgy, 2016, 13(4):489-497.
[11] 胡超晔. 基于Cassandra数据库集群的高并发读写系统的分析和应用研究[D]. 上海:上海交通大学, 2013.

The Study of the Data Storage and Retrieval for the Massive Data of MUSER based on Cassandra*

Shi Yue[1], Wang Feng[1,2], Li Pengcheng[1], Liu Yingbo[1*]

(1. Kunming university of science and technology, Kunming, 650500
2. Yunnan Observatory of Chinese academy of sciences, Kunming, 650216)

Abstract: Mingantu Ultrawide Spectral Radioheliograph produces massive observational data every day. In order to solve the problem of high latency of reading and writing, limited performance of capacity expansion and weak usability when the data is stored and managed by traditional relational database, The paper has carried out research on the application of MUSER data storage based on NoSQL. First of all, deeply analyze the characteristics of MUSER data, the requirement of storage and problems being solved in detail. Secondly, the MUSER data model is given as well as the MUSER column non-relational data model. The metadata and data are synchronized in NoSQL based on Cassandra. The MUSER massive astronomical data storage management system (MBDMS) is realized based on Cassandra. The experiment validates the efficiency, expansibility and feasibility of data storage and retrieval of the system. As a result, MBDMS can well meet MUSER data management needs, is an effective program to solve the current MUSER data storage problem.

Key words: non-relational database; storage; searching;